KLE Society's

KLE Technological University



**Exploratory Data Analysis**

**(21ECSC210)**

**Course Project Report on**

**“Leading Industries”**

*Submitted To*

**Mr. Karibasappa K. G Sir**

*Submitted By*

*Ashlesha Khanapure*

**Semester: 4th Div: C**

**2022-23**

**SCHOOL OF COMPUTER SCIENCE & ENGINEERING,**

KLE TECHNOLOGICAL UNIVERSITY

**Abstract**

The leading industries dataset is a collection of data that tracks the performance of different industries over time. This data can be used to identify emerging industries, track the growth of existing industries, and understand the factors that drive industry success.

The report concludes that the leading industries dataset is a valuable resource for businesses that are looking to identify new opportunities for growth and to make strategic decisions about their products and services.

**Table of Contents:**

**1. Introduction ………………………………………………… 05**

1.1 Overview of EDA Project ……………………………………………………………..

1.2 Importance of EDA Analysis …………………………………………………….....

1.3 Objectives ……………………………………………………………………………………

**2. Data Collection …………………………………………… 10**

2.1 Description of Data ………………………………………………………………………

2.2 Data collection process and sources …………………………………………….

2.3 Data processing steps performed pre- ……………………………………………

**3. Data Pre-processing ………………………………………….. 11**

3.1 Data pre-processing steps performed……………………..

**4. Data Exploration …………………………………………. 12**

3.1 Descriptive statistics of our dataset ………………………………………………

3.2 Visualizing data through plots, graphs, and charts …………………………

3.3 Identifying patterns, trends, or outliers in the data ……………………….

**5. Data Cleaning ………………………………………………. 18**

4.1 Steps to clean the dataset ……………………………………………………………

4.2 Addressing missing data and handling outliers …………………………………

4.3 Data transformation or normalization techniques used ………………

**6. Feature Selection …………………………………………. 18**

5.1 Process of selecting relevant features ……………………………………………

5.2 Dimensionality reduction techniques applied …………………………………

5.3 Rationale behind the feature selection process ……………………………

**7. Data Analysis ……………………………………………… 19**

**8. Insights and Findings ………………………………….. 19**

**9. Recommendations …………………………………….. 20**

**10. Conclusion …………………………………………………. 20**

**11. References ……………………………………………….. 20**

**Chapter-1**

**Introduction**

Overview of EDA Project

In this exploratory data analysis (EDA) project, we analyse a leading industries dataset to gain insights into their performance and trends. By examining variables such as revenue, profit, market share, and growth rate, we aim to understand industry dynamics. Our objectives include providing meaningful visualizations, summary statistics, and key insights that can guide decision-making processes and offer recommendations for the industries analysed. Through data cleaning, descriptive statistics, and data visualization, we aim to provide actionable insights and recommendations for decision-making in these industries

- **Data Understanding:** The first step is to familiarize ourselves with the dataset's structure and content. This involves understanding the columns, data types, and any relevant metadata. We should also clarify the criteria on which the industries are ranked to ensure a clear understanding of the data.

- **Data Cleaning:** Before conducting any analysis, it is essential to address any data quality issues. This includes handling missing values, removing duplicates, and addressing outliers that could potentially impact the accuracy of the insights.

- **Exploratory Data Analysis (EDA):** Perform exploratory data analysis to gain initial insights into the dataset. This includes generating summary statistics, such as mean, median, standard deviation, and quartiles, to understand the distribution and central tendencies of the industry rankings. Visualizations like bar charts, line charts, scatter plots, and histograms can be used to identify patterns and trends among the industries.

- **Ranking Analysis:** Analyse the ranking data to identify top-performing industries and industries that may need improvement. This could involve looking for correlations between different attributes and the industry rankings to understand which factors contribute most to high or low rankings.

**- Industry Comparison:** Compare the performance of different industries across various metrics. Identify strengths and weaknesses in each industry and how they relate to their rankings.

- **Predictive Analysis:** If the dataset includes historical data, consider performing predictive analysis to forecast future industry rankings. This could involve using time-series analysis or machine learning algorithms to predict future trends and rankings.

- **Insights and Recommendations:** Based on the analysis, draw meaningful insights from the data and develop data-driven recommendations for decision-making processes. These recommendations should address areas of improvement for industries with lower rankings and suggest strategies to maintain or enhance the success of top-performing industries.

- **Visualizations and Reports:** Present the findings, insights, and recommendations in a visually appealing and easy-to-understand format. Use graphs, charts, and concise summaries to communicate the results effectively to stakeholders.

Importance of EDA in data analysis

In the context of analysing a dataset of leading industries to provide insights and recommendations for decision-making processes, the importance of EDA (Exploratory Data Analysis) cannot be overstated.

EDA is a crucial step in the data analysis process, and its significance becomes even more evident in complex and large datasets like those from leading industries.

* **Understanding Data Composition:** EDA helps in comprehending the structure and composition of the dataset. It allows you to identify the different variables, their data types, and the overall data distribution, providing a clear picture of the information available for analysis.
* **Data Quality Assessment:** Before making any strategic decisions based on the dataset, it is vital to assess data quality. EDA aids in detecting missing values, outliers, inconsistencies, and other data quality issues, allowing you to address them before proceeding with the analysis.
* **Identifying Trends and Patterns:** Exploring the dataset through various visualization techniques helps uncover trends, patterns, and correlations among variables. Understanding these relationships can provide valuable insights into industry dynamics, customer behaviour, or market trends.
* **Feature Selection and Engineering:** In an industry dataset, there might be numerous features available. EDA can help identify the most relevant and informative features that can influence decision-making processes and improve the accuracy of predictive models.
* **Benchmarking and Performance Comparison:** EDA allows you to compare performance metrics of different industries or companies within the dataset. This analysis can highlight best practices and areas where improvement is needed.
* **Risk Assessment and Mitigation:** By exploring historical data and identifying patterns related to risk, EDA can help in risk assessment and formulating strategies to mitigate potential risks in specific industries.
* **Market Insights:** EDA can offer valuable insights into market dynamics, competitive landscape, and emerging trends. These insights can be leveraged to make informed decisions about market positioning and product offerings.
* **Identifying Business Opportunities:** Analysing leading industries' data can reveal untapped opportunities or niche markets that decision-makers can exploit to drive growth and innovation.
* **Data-Driven Decision Making**: EDA provides a strong foundation for data-driven decision-making. The insights gained from EDA can support decision-makers in formulating evidence-based strategies and actions.
* **Effective Communication:** Visualizations and summary statistics from EDA facilitate effective communication of findings to stakeholders. It aids in presenting complex information in a more understandable and compelling manner.
* **Iterative Analysis:** EDA often involves an iterative process of exploration, where new insights lead to more targeted questions and further analysis. This iterative nature ensures a thorough understanding of the data and maximizes the potential for valuable insights.

By conducting thorough EDA, data analysts can lay a solid foundation for subsequent data analyses and model building, contributing to the overall success of data-driven projects.

**Objectives**

* Clean and prepare data for analysis.
* Explore and visualize trends and patterns.
* Analyze industry performance.
* Identify leading and lagging Industries.
* Investigate correlations between variables.
* Provide insights and recommendations.
* Present findings in a clear and concise manner.
* Inform decision-making processes.

**Chapter-2**

**Data Collection**

Description of the data

Figure : Column with not-null count and data types

Data collection process and sources

Our first aim was to find a good set of data to work upon an continue our Exploratory Data Analysis (EDA).

- The dataset used in the case study:

● Dataset has 321 rows and 10 columns.

● All values are non-null.

● Dataset Shape: (321, 10).

○321: Rows

○10: Columns (Attributes)

● CSV File Size: Leading\_Industries.csv (32.8KB)

● Data Types: int64(2), float64(3), object(5).

- Data Loading: Import the leading industries dataset from a specified source (e.g., CSV file, database). Load the dataset into the system for further analysis.

LINK- <https://www.kaggle.com/datasets/earljohnmasaga/leading-industries-dataset>

**Chapter-3**

**Data pre-processing**

Data pre-processing steps

* Checking for NULL values.
* Checking for duplicate tuple and removing them.

Data pre-processing steps performed

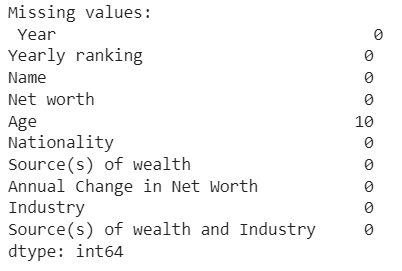


Figure 2 : Column with missing values.

**Chapter-4**

**Data exploration**

Descriptive statistics of our dataset

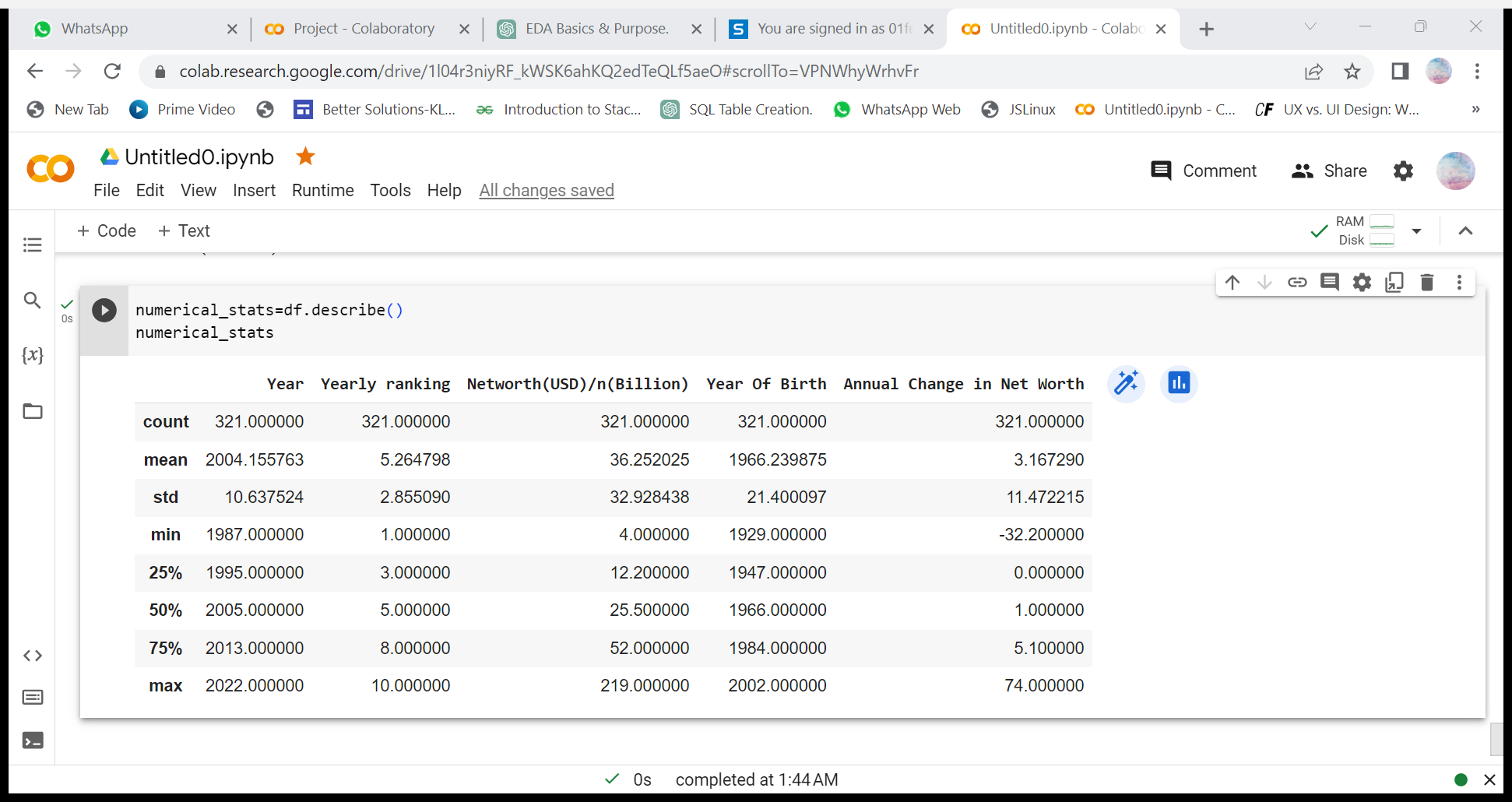
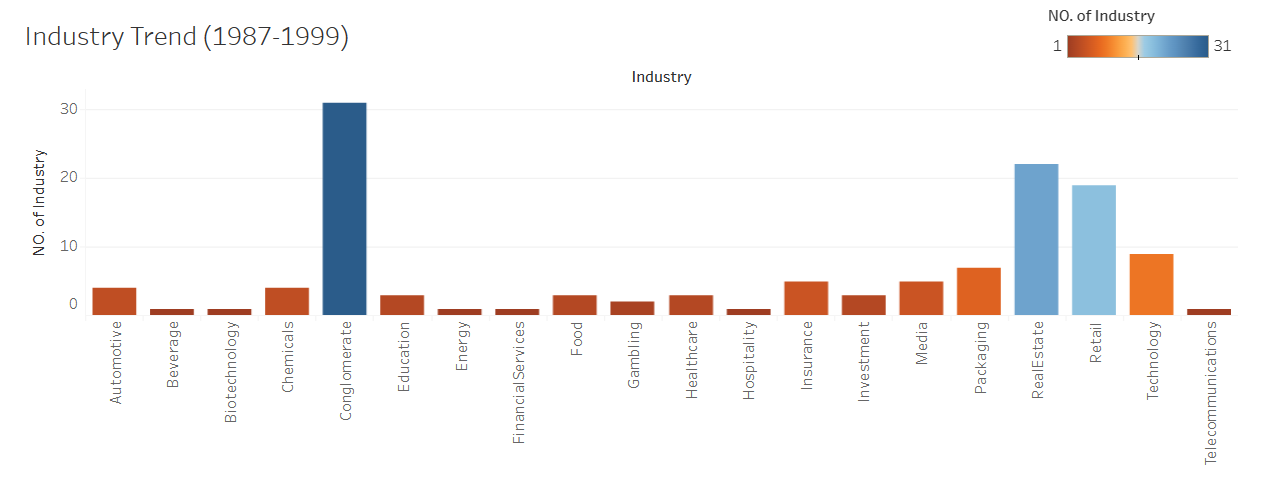


Figure : Descriptive statistics of important numeric column

**Data Visualisation**

1. Industry Trend (1987-1999)

Figure 4: Histogram for industry trend in year (1987-1999)

* Conglomerate had the highest number of industries.
* Real Estate is second and Retail is third.

2.Industry Trend (2000-2022)

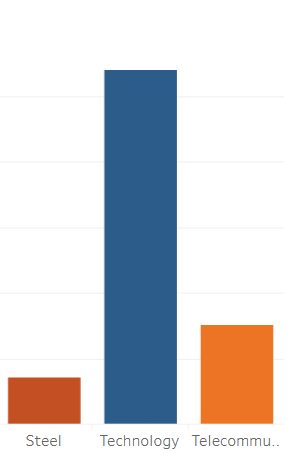
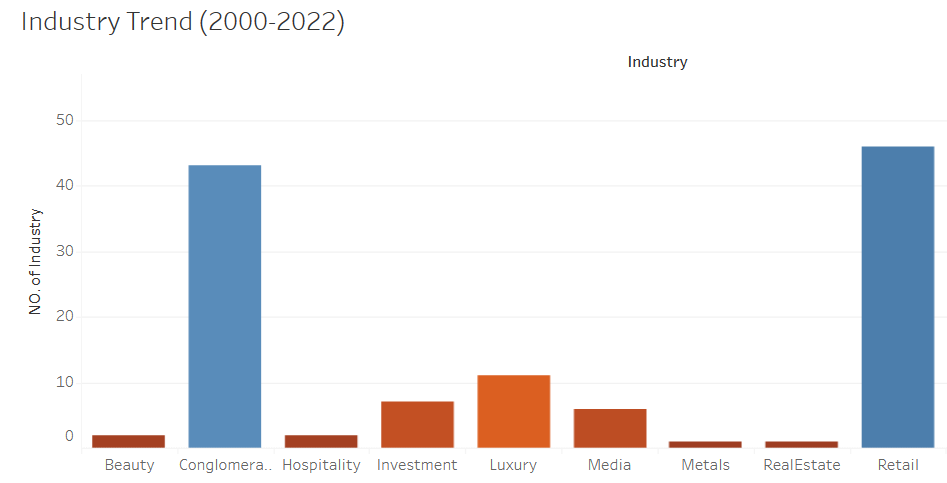


Figure 5: Histogram for industry trend in year (2000-2022)

* Technology had the highest number of industries.
* Retail is second and Conglomerate is third.

3.The World’s Billionaires-1987

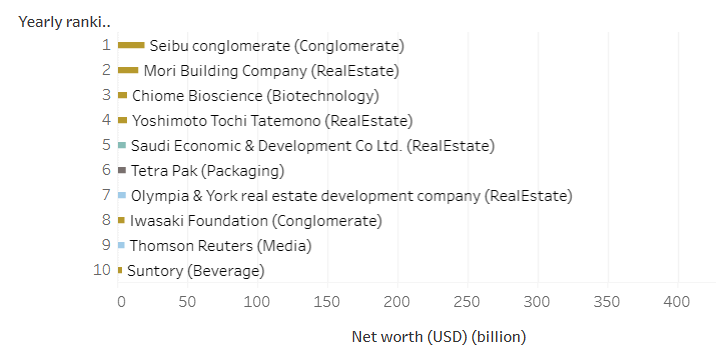


Figure 6: Gantt chart of world’s billionaires in 1987.

* Seibu Conglomerate had the highest net worth in 1987.

4.The World’s Billionaires-2020

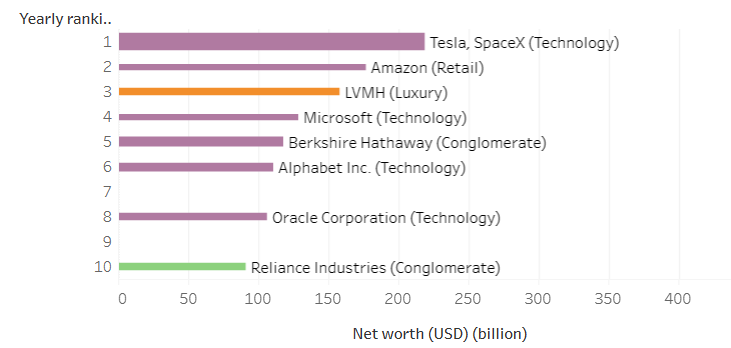


Figure 7: Gantt chart of world’s billionaires in 2022.

* Tesla, Space X had the highest net worth in 2022.

5. Wealth Distribution by Industry

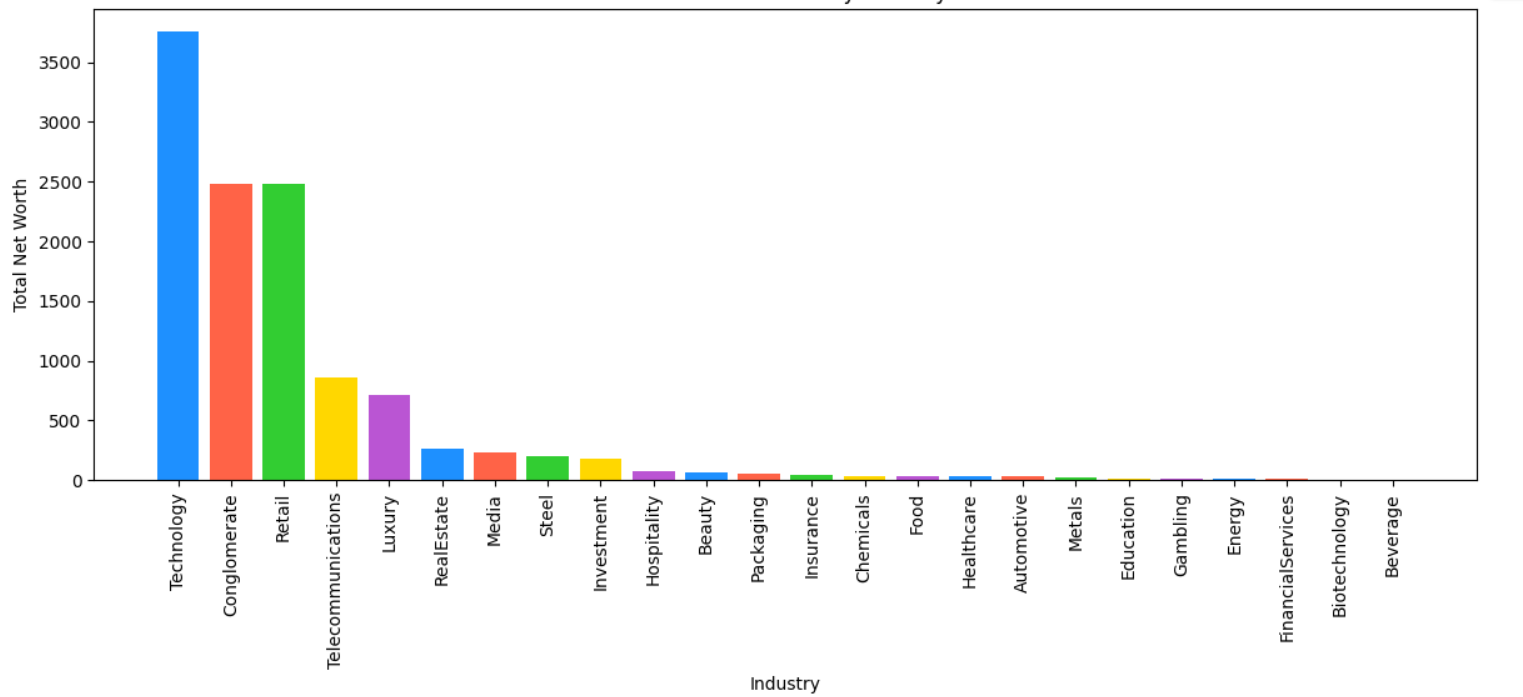


Figure 8: Histogram for wealth distribution by industry.

* Technology is the dominant Industry with highest net worth.

6. Relationship between Age and Net Worth

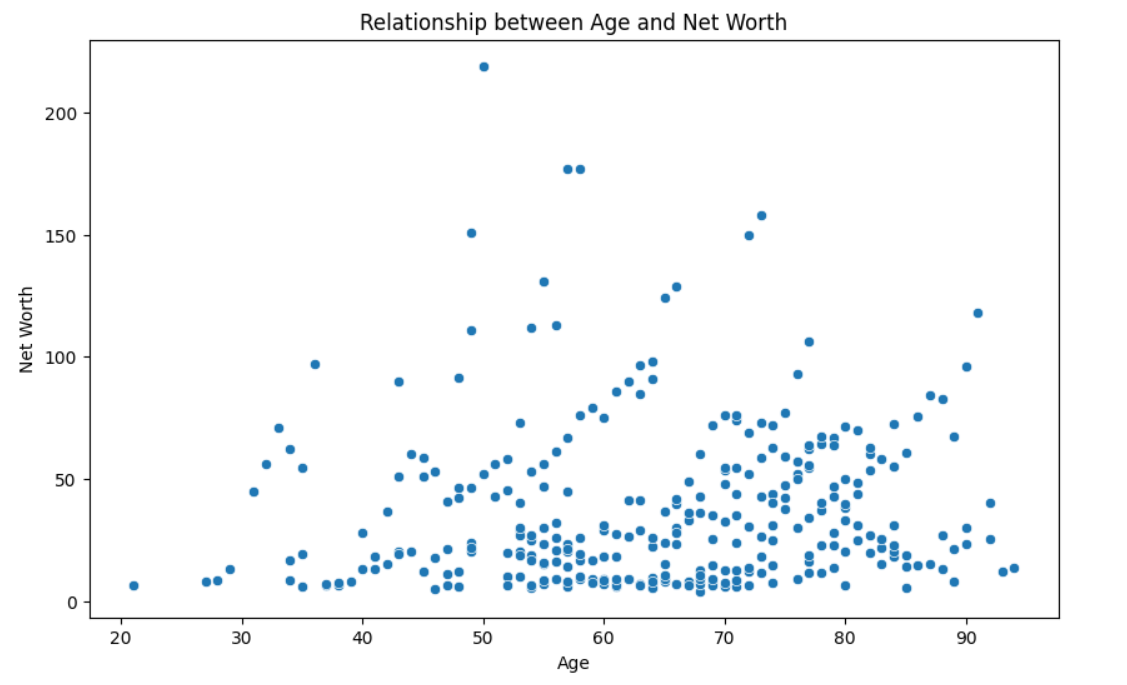


Figure 9: Scatter-plot showing correlation between age and net worth.

* Between Age group 50-70, Net worth is high.

7. Heat map of Net Worth

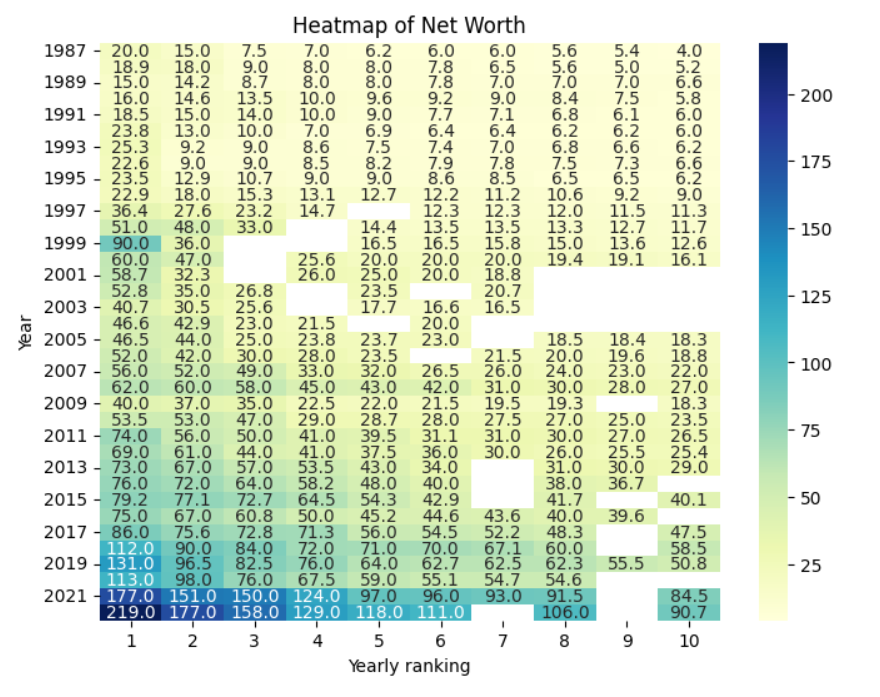


Figure 10: Heat-map of net worth.

* 2022 had more Yearly rankings.

8. Nationality Distribution

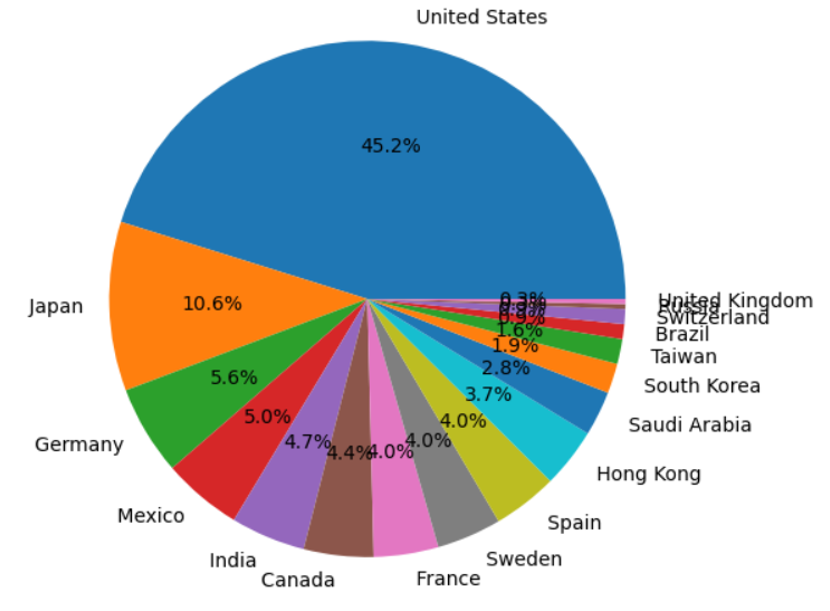


Figure 11: Pie chart of nationality distribution.

* United States has the highest Nationality Distribution.

9.Relationship between Year, Networth, and Annual Change in NetWorth

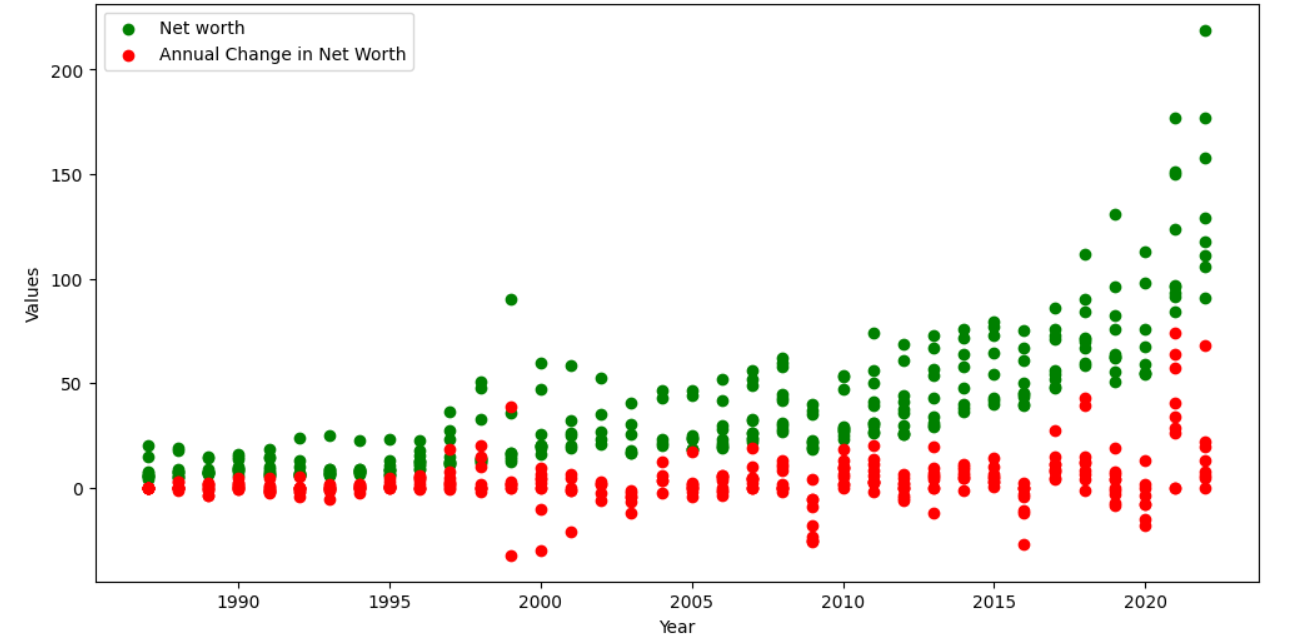


Figure 12: scatter-plot for increase in net worth.

* Net Worth increased with Increase in year.
* Annual change in Net worth was almost 0.

**Chapter-5**

**Data Cleaning**

1. Removing of NULL values-

df.dropna()

2. No duplicates found in the dataset.

Addressing missing data and handling outliers

Missing data / null values were observed by looking at the dataset and filled as according to the significance of the attributes.

Outliers were not present in the dataset so there was no need of handling it.

**Chapter-6**

**Feature Selection**

Feature selection is a critical step in the data analysis process, especially when analysing a dataset of leading industries to provide insights and recommendations for decision-making processes. The selected features for this project include revenue, profit, market share, growth rate, expenses, industry type, geographic region, and company size. These variables will be analysed and visualized to gain insights into industry performance and trends, leading to actionable recommendations for decision-making.

**Chapter-7**

**Data Analysis**

Data analysis is an iterative process, and it is essential to document all steps taken during the analysis for reproducibility and transparency. Data analysis is a crucial step in the process of analysing a dataset of leading industries to provide insights and recommendations for decision-making processes. We will clean the data and perform descriptive statistics. It involves performing statistical analysis, applying appropriate analytical methods and models, and interpreting the results to draw meaningful conclusions.

Through data visualization and industry comparisons, we'll identify trends, correlations, and regional insights. The key findings will drive actionable recommendations for decision-making in the leading industries.

**Chapter-8**

**Insights and Findings**

Year, Yearly ranking, Net worth, Source of wealth and Annual change in net worth were used find out the leading industries.

Naive Bayes algorithm demonstrated good predictive capabilities and was suitable for this classification task.

**Chapter-9**

**Recommendations**

This help to get information about leading industries and which branch of industries are in top. This analysis also helps people to invest in what type of industries, and also this is helpful for knowing the growth of the country.

**Chapter-10**

**Conclusion**

From the above Analysis we conclude that Technology is the leading Industry. Investing in Technology industry would be profitable. Beverage is a lagging Industry. Between Age group 50-70, Net worth is high. United States has the highest Nationality Distribution. Conglomerate had the highest number of industries.

**Chapter-11**

**References**

* + https://www.kaggle.com/datasets/earljohnmasaga/leading-industries-dataset
  + "Exploratory Data Analysis" by John W. Tukey Addison-Wesley, 1977. ISBN: 978-0201076165

**Acknowledgments**

Mr. Karibasappa K G: Thank you for your guidance, support, and valuable insights throughout the course. Your expertise and dedication have been instrumental in shaping our understanding of Exploratory Data Analysis (EDA) techniques.

Ms. Shilpa: I am thankful for your continuous encouragement and feedback. Your enthusiasm for data analysis has been inspiring, and your constructive comments have helped us improve our work.

**THANK YOU**